Introducing epistasis and high performance computing into stepwise model selection for GWAS

Alexander E. Lipka
Assistant Professor of Biometry
Department of Crop Sciences
UIUC

Genome-wide association study (GWAS)
Association with $\alpha$-Tocopherol Levels in Maize Grain

Peak SNP is within ZmVTE4

- Identify genomic regions associated with a phenotype
- Fit a statistical model at each SNP in genome
- Use fitted models to test $H_0$: No association with SNP and phenotype

Unified mixed linear model: widely used in plant GWAS

\[ Y_i = \mu + \sum \beta_j C_j + \alpha + (\text{Line}) + \epsilon_i \]

- $K = \text{kinship matrix}$
- $\epsilon_i \sim i.i.d. \ N(0, \sigma^2)$

Drawback: Tests one marker at a time
Solution: Stepwise model selection

I helped implement this approach into the TASSEL 5 GUI:
- NAM panel: Markers nested within families
- Diversity panel: Markers not nested within families

Epistasis is the interaction between loci

- B locus controls black vs. brown color
- E locus controls yellow vs. not yellow color
- Dogs with “ee” genotype are yellow, regardless of genotype at B locus

Modeling epistasis as two-way interaction terms

\[ Y_i = \mu + \sum_{j \in I} \beta_j C_j + \sum_{(u,v) \in U} \gamma_{uv} X_{ul} X_{vl} + \epsilon_i \]

- $I$ is a subset of markers with additive effects in model
- $U$ is a subset of markers with two-way epistatic effects in model
- Determining the optimal model:
  - AIC, BIC, mBIC
  - Permutation procedure

Yu et al. (2006)

Yu et al. (2006)

Yu et al. (2006)

Lipka et al. (2013)
Implementation of stepwise epistatic model selection

- Incorporated the search for two-way epistatic effects into modified TASSEL stepwise model selection module

Goal - to make search for epistasis more widespread:
- Program freely available to public
- Used within the popular TASSEL framework

It may take years to run stepwise epistatic model selection on a single thread

<table>
<thead>
<tr>
<th>Data</th>
<th>Individuals</th>
<th>SNPs</th>
<th>Two-way SNP interactions</th>
<th>Single-threaded walltime on commodity laptop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human eGWAS</td>
<td>300</td>
<td>3,000,000</td>
<td>~60 billion</td>
<td>~6 hours for SNPs</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>~1,056 years for SNP interactions</td>
</tr>
</tbody>
</table>

Solution: implement multithreading
- Should result in reduced computational time

Multithreaded stepwise model selection was assessed in ten US-NAM families

Settings used to assess performance:
- Entry P-value threshold: 1.0x10^-9
- Exit P-value threshold: 2.0x10^-9
- Markers were nested within family
- Enabled a “worst-case scenario” assessment of computational time

Multithreading decreased 61% walltime

Preliminary results (putatively) identify two pairs of epistatic carotenoid genes

Carotenoid biosynthetic pathway

Orange kernel color

Scalability analysis for epistatic model
Projected outcome looks positive

Future work: Simulation studies

- Assess the ability to detect epistatic effects:
  - Under various genetic architectures
  - Under various model selection criterion
  - Under various allele frequencies
  - Under various sample sizes and marker densities
  - Consider additive x dominance, dominance x additive, and dominance x dominance epistasis
  - NAM design vs. a typical diversity panel

<table>
<thead>
<tr>
<th>Data</th>
<th>Single-threaded walltime on commodity laptop</th>
<th>Projected 96-threaded walltime on a 96 core server</th>
</tr>
</thead>
<tbody>
<tr>
<td>US Maize NAM [no epistasis]</td>
<td>~3 hours</td>
<td>~1 minute for SNPs</td>
</tr>
<tr>
<td>US Maize NAM [epistasis]</td>
<td>~1.6 years</td>
<td>A few weeks</td>
</tr>
</tbody>
</table>

Future work: Long-term goals

- Investigate other subset selection approaches, in particular the LASSO:
  - Assess power and computational speed
- Migrate code to lower-level programming language:
  - C and/or C++ are the target languages
- Perform analysis on human data:
  - Efficient analysis of larger human data sets presents a challenge

Final Thoughts

- Searching for epistatic signals is difficult:
  - Rare allele frequencies
  - Inherent computational burden
  - Linkage disequilibrium between markers and causal variants
  - Multiple testing correction
- Nevertheless searching for epistasis is important
- The program we are developing could facilitate the search for epistasis:
  - Uses a statistical model that simultaneously accounts for multiple loci
  - User-friendly interface
  - Freely available to the research community

Acknowledgements

Nested Association Mapping (NAM) panel is used for analysis

Ideal for dissecting complex traits:

- Inflorescence
- Flowering time
- Leaf architecture
- Stalk strength
- Day length
- Kernel
- Composition
- Leaf Blight
TASSEL5 has a significant body of followers

- Written in Java
- Incorporates stepwise model selection

Modified Bayesian Information Criterion (mBIC)

- Select the model that minimizes the following criterion:

\[ n \log(\text{RSS}) + (p + q)n \log n + 2p \times \log \left( \frac{N_m}{2} - 1 \right) + 2q \times \log \left( \frac{N_e}{2} - 1 \right) \]

- Ten NAM families with kernel color ranging from light yellow to dark orange
- Sample size: 1,555 NAM RILs
- Number of Markers: 1,106

Results: Stepwise model selection for orange kernel color